Original Research

Item-Level Assessment of the Irritable Bowel Syndrome Quality of Life Questionnaire in Patients With Diarrheal Irritable Bowel Syndrome

David A. Andrae, PhD[1]; Paul S. Covington, MD[1]; and Donald L. Patrick, PhD, MSPH[2]

[1]Furiex Pharmaceuticals Inc, Morrisville, North Carolina; and [2]Department of Health Services University of Washington, Seattle, Washington

ABSTRACT

Background: In previous studies, the Irritable Bowel Syndrome Quality of Life (IBS-QOL) instrument has been determined to have good measurement properties for general irritable bowel syndrome (IBS) and the diarrheal IBS (IBS-d) subtype in clinical trials.

Objective: This article aims to extend the true-score analyses that have been previously conducted to evaluate the IBS-QOL in IBS-d patients.

Methods: Item response theory analysis was conducted by fitting models to responses from 753 patients with severe IBS-d from a recent clinical trial. Three item response theory models, the constrained graded response model (CGRM), the unconstrained GRM (UGRM), and the testlet response model (TRM), were fit to the 34 items of the IBS-QOL questionnaire. Subsequently, differential item functioning (DIF) for patient sex was assessed by fitting nested models by applying likelihood ratio tests. Model latent trait estimates were then compared with the IBS-QOL score and the IBS Symptom Severity Score.

Results: Model fits improved with complexity, with the TRM model fitting best compared with the CGRM and UGRM. The DIF evaluation for patient sex flagged 17 items for the CGRM and 9 items for the UGRM; however, these items were not found to have much effect on the overall estimation of the latent trait. Differential testlet functioning was not indicated, and no items exhibited potential DIF under the TRM because likelihood ratio tests were not statistically significant. Comparison of latent trait estimates to the IBS Symptom Severity Score and IBS-QOL questionnaire revealed high Spearman correlations (0.47 and $\geq$0.99, respectively).

Conclusion: Previous true-score approach results were supported by the IBS-QOL item-level analysis. Further, the IBS-QOL total score was found to be a valid measure of perceived quality of life for IBS-d patients when compared with more sophisticated model-based estimates of perceived quality of life. ClinicalTrials.gov identifier: NCT01130272. (*Clin Ther*. 2014;36:663–679) © 2014 The Authors. Published by Elsevier HS Journals, Inc. All rights reserved.

Key words: diarrhea, differential item functioning, graded response model, IBS-QOL, irritable bowel syndrome, item response theory, patient-reported outcomes, psychometrics, quality of life.

INTRODUCTION

Irritable bowel syndrome (IBS) is a debilitating functional gastrointestinal disorder that affects an estimated 10% to 15% of people in Western cultures.[1] Irritable bowel syndrome is characterized by recurrent abdominal discomfort and pain associated with altered bowel habits.[2] Current medical guidelines (ie, Rome III criteria) determine IBS subtypes by stool consistency patterns, such as diarrheal IBS (IBS-d), IBS with constipation, or IBS with both constipation and

Accepted for publication April 9, 2014.
http://dx.doi.org/10.1016/j.clinthera.2014.04.009
0149-2918/$ - see front matter

diarrhea.[3] In addition, IBS can negatively affect an individual's perceived quality of life (QOL), and research indicates that patients with IBS incur significant direct and indirect costs associated with the condition. Current tolerable and effective pharmacologic treatments for IBS are limited, with current treatment options that include antispasmodics, antidepressants, antidiarrheal agents, and alosetron.[4]

Recently, a large Phase II clinical trial of eluxadoline, a mixed μ-opioid agonist and δ-opioid antagonist in patients with IBS-d was reported.[5] One of the exploratory objectives of the trial was to evaluate the psychometric properties of perceived QOL instruments administered to the patients enrolled in the study. Previous psychometric validation has focused on a true-score theory approach (ie, classic test theory [CTT]) to the IBS-QOL with the original CTT validation work performed on an overall, non-subtyped patient set.[6–8] The IBS-QOL questionnaire is a 34-item instrument designed from a needs-based conceptual paradigm, and scoring is composed of a total score and 8 subscale scores that have been validated for a general IBS patient population. Recent research on the true-score properties of the IBS-QOL in IBS-d–specific patients has yielded information about the total score and not the substructure of the IBS-QOL is the best measurement model for IBS-d patients.[9] Specifically, Andrae et al[9] concluded that the IBS-QOL may have individual items that do not pertain directly to IBS-d patients but that the questionnaire is likely measuring a unidimensional construct. Further, their analyses did not directly support the original substructure as determined by the original research on a set of patients with IBS with both constipation and diarrhea. Because the IBS-QOL questionnaire has been widely studied via CTT methods, we intended to further explore these recent developments with the use of item-centered methods. This article uses several item response theory (IRT) approaches to extend the previous psychometric assessments of the IBS-QOL questionnaire in an attempt to determine whether item-level data elucidate aspects of measurement of perceived QOL beyond previously validated scoring methods.

At the heart of CTT approaches to psychometrics is the idea that the latent construct of interest is best measured by a single score from the instrument (eg, the sum of scored items or, as in the IBS-QOL questionnaire, a standardized total score). Such an approach typically assumes that all items contribute equally to measuring the construct. The IRT methods, in contrast, focus evaluation on the individual items themselves by fitting model parameters that relate items to the latent trait under study. The focus on individual items in the IRT methods is ideal for assessing the IBS-QOL questionnaire in IBS-d patients because there are previous indications that certain items do not perform well in using the IBS-QOL questionnaire to assess perceived QOL for IBS-d patients.

The current paper aims to use IRT modeling to complement and extend the results of the classic psychometric assessments described by Andrae et al.[9] Specifically, different graded response models (GRMs)[10] were fit to the items of the IBS-QOL questionnaire. Although the CTT methods focus on the characteristics of the instrument as a whole (ie, that the sum or total score from an instrument serves as a best estimate of some latent trait of interest), modern methods use individual item responses as means of assessing latent traits. The idea is that individual items each measure some aspect of a latent trait that is not directly observable. Therefore, by directly modeling the contributions of the items to the measurement of this latent trait, one can gain a richness of measurement not gotten by summed scores. In turn, then, by modeling IBS-QOL questionnaire items as opposed to overall scores, it is hoped that better precision can be gained with regard to characterizing and measuring perceived QOL—the latent trait in the current case—for IBS-d patients.

METHODS

All analyses were applied to a sample of 753 patients who completed a baseline administration of the IBS-QOL questionnaire while participating in a Phase II study of eluxadoline for the treatment of IBS-d. Patients were recruited between May 2010 and April 2011 from 292 centers in the United States. Further details of study conduct are reported elsewhere.[5] Patients were diagnosed as having IBS-d according to Rome III criteria. The study was completed in compliance with good clinical practice guidelines and the Declaration of Helsinki. Analyses were conducted using R version 3.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

IRT Models

As stated above, previous work on the IBS-QOL has centered on the CTT methods. In a true-score approach, some measure of overall instrument score is the fungible unit by which an instrument measures the latent construct under study. The IRT methods shift the fungible unit to individual item responses. These responses are used jointly to estimate the latent construct. In the current case, the latent construct is IBS-d–related QOL, represented by the Greek letter θ, and because items are rated by patients on 5-point Likert-type responses, analyses were performed using models appropriate for graded responses.[10] Two types of GRMs were fit to the data.

The first model fit was the constrained GRM (CGRM). In this model, only item intercepts are estimated, not slopes. For patient i and item j, the probability of response u_{ij} is greater than or equal to a response level k, as given in model 1:

$$\Pr(u_{jk} \geq k|\theta) = \frac{\exp\,(\alpha_{jk}+\theta_i)}{1+\exp\,(\alpha_{jk}+\theta_i)} \tag{1}$$

where α_{jk} represent the thresholds between ordinal responses (ie, the location parameters), and θ is the latent construct measured by the items. In model 1 the linear predictor for each item is constrained to be equal to 1 and is constant across items. Probabilities of actual responses are then calculated by the differences between adjacent categories by taking the differences between response categories:

$$\Pr\,(u_{ij} = k|\theta) = \Pr(u_{ij} \geq k|\theta) - \Pr(u_{ij} \geq k+1|\theta)$$

where $\Pr(u_{ij} \geq 1) \equiv 1$.[11] Another way of conceptualizing the CGRM is that each item is given equal weight in measuring θ and that the equal weighting is standardized to a value of 1, indicating that a patient's perceived QOL is, ostensibly, measured only by the item intercepts (ie, the location of the items on the latent continuum). The CGRM fit in the current article is similar in parameterization of the linear predictor to the partial credit model, a member of the Rasch family of IRT models.[12] Because the Rasch family of models define response probabilities not in terms of differences between adjacent categories but rather in terms of the proportion of total response probability for each item (ie, the so-called divide by total models),[13] the CGRM was fit here for consistency of the statistical paradigm.

Although the CGRM is a parsimonious means of relating item responses to θ, such an approach may not be reasonable for all item sets. Therefore, an unconstrained GRM (UGRM) was also fit to see whether the CGRM assumptions were reasonable for the IBS-QOL data as in model 2:

$$\Pr(u_{ij} \geq k|\theta) = \frac{\exp\,(\alpha_{jk}+\beta_j\theta_i)}{1+\exp\,(\alpha_{jk}+\beta_j\theta_i)} \tag{2}$$

which allows the item slopes β_j to vary across the j items of a given scale and is also estimated so that it can be a value different than 1. The slope parameters are interpreted as higher value that represented higher discrimination between levels of θ. Comparing model 2 to model 1, one can see many similarities. In fact, the only difference is that the βj values in model 1 are all set to equal rather than estimated for each item as in model 2. Likewise, the intercept terms α_{jk} in both models 1 and 2 can be thought of as the location of a given response k to a given item j on the latent variable scale. In the case of graded responses, higher slope values mean that individual response probabilities are more well defined and distinct from one another across responses than for items with lower slope values and item intercepts. Thus, UGRMs with slopes that span a large range indicate that discrimination can be had over a wide range of latent variable values.

Because it is hypothesized that the IBS-QOL questionnaire items will not equally relate to θ for an IBS-d patient set because the items were generated for a patient population with nonsubtyped IBS, it is expected that the UGRM fit the data better than the CGRM. This hypothesis is supported by the previous factor analyses of the IBS-QOL questionnaire because the factor loadings in the single-factor model were found to be unequal.[9] Determination of relative model fit is assessed by a significant likelihood ratio (LR) χ^2 (ie, by comparing −2 times the difference in the log likelihood of the models to a χ^2 distribution with degrees of freedom equal to the difference in the number of model parameters).

Testlet Response Model

Like many other IRT models, GRMs have 2 core assumptions: (1) the unidimensionality assumption (the latent construct being measured is homogeneous) and (2) the local independence assumption (that the individual items are independent from one another when accounting for θ). One of the criticisms of IRT models that allow estimation of different slopes is that such different discrimination parameters across items could suggest a lack of unidimensionality of the latent construct being measured, thereby violating assumption 1.[14] Questions can also be raised as to whether so-called local dependencies negatively influence model assumptions because differential slopes could suggest interrelationships among items, potentially violating assumption 2.[14] Because previous CTT approaches to psychometric analyses of the IBS-QOL questionnaire revealed indications of minor deviations from the single-factor model for the items, indicating that the unidimensionality and local independence assumptions for the GRM may not hold, an alternative conceptualization was warranted.[9] Testlet response theory (TRT) approaches the psychometric assessment of an instrument using groups of items, or testlets, as the fungible unit of measurement of latent traits.[14–16] Comparatively, the fungible units of measurement in true-score theory and IRT are, respectively, the total score from the instrument and the items. Although TRT is usually applied as an a priori test construction method (eg, several comprehension questions about a single reading passage), in the current case we are attempting to see whether item content similarities may lead to the deviations in the single-factor structure and thus manifest as potential violations of the conditional independence of the GRM. By fitting a testlet response model (TRM) one can perform statistical evaluation of the effect due to potential departures from the GRM assumptions because item local dependence patterns can be built into the model. To account for potential local dependencies between sets of items, random effects will be fit in addition to the GRM in models 1 and 2. Specifically, we define the TRM as follows:

$$\Pr(u_{ij} \geq k|\theta) = \frac{\exp\,(\alpha_{jk}+\beta_j\theta_i+\zeta_{it(j)})}{1+\exp\,(\alpha_{jk}+\beta_j\theta_i+\zeta_{it(j)})} \qquad (3)$$

where $\zeta_{it(j)} \sim \text{Normal}(0,\text{var}(\zeta_{it(j)}))$ represents the crossed random effects of testlet t containing specific items j. Model 3 is a special case of the bifactor model,[17] with specific subsets of items loading on the nongeneral factors (ie, testlets). For identification of the model, the slopes of each item on each testlet are constrained to be equal to that same item's slope on the general factor. This allows the variance within a testlet, var($\zeta_{it(j)}$), to be estimated across these slopes, thereby accounting for any dependencies among items within a testlet. By comparison, the traditional bifactor model constrains variances of all factors to be 1, allowing estimation of item slopes to vary between general and specific factors. Therefore, the TRM model, to the extent the model is properly constructed, statistically controls for local dependencies by allowing estimation of variances across items within testlets,

Table I. Items grouped into testlets for the testlet response model.

Testlet No.	Items
I	11. I have to watch the amount of food I eat because of my bowel problems. 23. I have to watch the kind of food I eat because of my bowel problems. 28. I feel frustrated that I cannot eat when I want because of my bowel problems.
II	12. Because of my bowel problems, sexual activity is difficult for me. 20. My bowel problems reduce my sexual desire.

In the current case, 2 testlets were determined by content review. Table I displays the items that were grouped into 2 testlets, roughly relating to "food" and "sexual" items. Although other testlet configurations may result in better identification of fungible units within the IBS-QOL questionnaire, further research beyond the scope of the current article, including qualitative assessment by patients and experts, is warranted to adequately determine the appropriate testlets.

Differential Functioning

Because previous investigations of IBS-d treatments have almost exclusively included only women, differential item functioning (DIF) due to patient sex was assessed by comparing the fit of models 1 and 2 on the IBS-QOL questionnaire with the following models.

$$\Pr(u_{ij} \geq k|\theta) = \frac{\exp\,(\alpha_{jk}+\theta_i+\gamma_j x+\delta_j\theta_i x)}{1+\exp\,(\alpha_{jk}+\theta_i+\gamma_j xx+\delta_j\theta x)} \qquad (4)$$

$$\Pr(u_{ij} \geq k|\theta) = \frac{\exp\,(\alpha_{jk}+\beta_j\theta_i+\gamma_j x+\delta_j\theta_i x)}{1+\exp\,(\alpha_{jk}+\beta_j\theta_i+\gamma_j xx+\delta_j\theta_i x)} \qquad (5)$$

where group effects for sex x and the sex × perceived QOL-level interaction effects are sequentially added to the model. Iterative model fitting and evaluation used the "multipleGroup" function within the MIRT package, version 1.1 (Vector Psychometric Group, Chapel Hill, North Carolina). Models 4 and 5 represent the DIF assessments for the CGRM and UGRM from above, respectively. The LR tests between models 1 and 4 and models 2 and 5 were evaluated as indications of overall presence of DIF for a given item as tested by a LR χ^2 test with 2 *df*. In an attempt to identify specific items that may exhibit DIF, we compared each model to all subset models in which single items were excluded in turn, and the 33 remaining items were fit to the GRM models. By this method, all included items serve as anchors to determine DIF of the excluded item. To adjust for the multiplicity of DIF assessments, LR χ^2 values were individually assessed for significance at the Bonferroni-adjusted α level. The Bonferroni adjustment was taken to account for the multiplicity of hypothesis tests for DIF.

Similarly, DIF for the TRM was also assessed by applying the methods above. In the case of the TRM, models 3 and 6 with the random effects for testlets $\zeta_{it(j)}$ were introduced to account for testlet effects, and tests of γ_j and δ_j were performed. Model 6 represents the TRM for DIF

$$\Pr(u_{ij} \geq k|\theta) = \frac{\exp\,(\alpha_{jk} + \beta_j\theta + \zeta_{it(j)} + \gamma_j x + \delta_j\theta_i x)}{1 + \exp\,(\alpha_{jk} + \beta_j\theta + \zeta_{it(j)} + \gamma_j x + \delta_j\theta_i x)} \tag{6}$$

Similarly to the DIF assessments for the item-level models, TRMs (ie, models 3 and 6) were compared across sexes. Further, submodels in which testlets were removed, in turn, were also fit to determine potential differential testlet functioning and DIF within the TRM. As was done for the GRM models above, all TRM functioning comparisons used a Bonferroni adjustment.

Associations between Models and Traditional Metrics

Finally, in an attempt to relate the estimation of the latent trait θ to traditional measures of IBS-d severity, factor scores from the CGRM, UGRM, and TRMs were compared with the IBS Symptom Severity Scale (IBS-SSS)[18] and the IBS-QOL total scores via Spearman correlation coefficients. The IBS-SSS scores range from 0 to 500, with higher scores indicating worse symptoms, and the IBS-QOL questionnaire scores range from 0 to 100, with higher scores representing better perceived QOL. Estimated θ values were all on the z-score metric, with higher values indicating worse perceived QOL.

One can further evaluate the CGRM, UGRM, and TRM models by viewing the model fit results from a factor-analytic approach where the GRM or TRM slopes are rescaled to be comparable to traditional factor analysis loadings (λ_j).[17,19] That is:

$$\lambda_j = \frac{\alpha_j/1.702}{\sqrt{1 + (\alpha_j/1.702)^2}} \tag{7}$$

where the factor 1.702 transforms the slopes to the normal ogive metric. Further, as a comparison of model estimation, factor scores and information functions for θ were computed for the CGRM, UGRM, and TRMs.

Table II. Demographic and baseline characteristics.

Characteristic	No. of Patients	Mean (SD)
Total IBS-QOL score	753	53.0 (21.13)
Total IBS-QOL score by sex		
Female	526	51.3 (21.18)
Male	227	57.2 (20.47)
Baseline IBS Symptom Severity Scale score	614	342.6 (64.38)
Baseline worst abdominal pain intensity (7-day mean, 0-10 rating)	753	5.89 (1.633)
Baseline stool consistency rating (7-day mean, Bristol Stool Scale)	753	6.22 (0.426)
Baseline No. of daily bowel movements (7-day mean)	753	4.81 (3.274)

IBS-QOL = Irritable Bowel Syndrome Quality of Life.

All models were fitted with the MIRT package, version 1.1, via R, version 3.02, using the Metropolis-Hastings-Robins-Monro (MH-RM) fitting algorithm.[19,20] The MH-RM algorithm was used instead of the usual expectation-maximization algorithm because the MH-RM has been found to work more efficiently for models

Table III. Item response frequencies.

Item[*]	Distribution of Item Responses, No. (%)				
	1	2	3	4	5
1	79 (10.5)	175 (23.2)	207 (27.5)	200 (26.6)	92 (12.2)
2	86 (11.4)	163 (21.7)	177 (23.5)	194 (25.8)	132 (17.6)
3	20 (2.7)	85 (11.3)	148 (19.7)	256 (34.0)	244 (32.4)
4	193 (25.6)	226 (30.0)	176 (23.4)	105 (13.9)	53 (7.0)
5	65 (8.6)	137 (18.2)	156 (20.7)	209 (27.8)	185 (24.6)
6	183 (24.3)	188 (25.0)	157 (20.9)	142 (18.9)	82 (10.9)
7	65 (8.6)	160 (21.2)	171 (22.7)	186 (24.7)	171 (22.7)
8	128 (17.0)	183 (24.3)	168 (22.3)	158 (21.0)	116 (15.4)
9	201 (26.7)	212 (28.2)	166 (22.1)	117 (15.6)	56 (7.4)
10	225 (29.9)	228 (30.3)	136 (18.1)	107 (14.2)	56 (7.4)
11	65 (8.6)	105 (14.0)	178 (23.7)	223 (29.7)	181 (24.1)
12	287 (38.2)	178 (23.7)	145 (19.3)	91 (12.1)	51 (6.8)
13	174 (23.1)	207 (27.5)	155 (20.6)	126 (16.7)	91 (12.1)
14	221 (29.4)	231 (30.7)	118 (15.7)	119 (15.8)	63 (8.4)
15	28 (3.7)	121 (16.1)	162 (21.5)	213 (28.3)	228 (30.3)
16	101 (13.4)	213 (28.3)	187 (24.9)	174 (23.1)	77 (10.2)
17	246 (32.7)	152 (20.2)	116 (15.4)	129 (17.2)	109 (14.5)
18	98 (13.0)	183 (24.3)	181 (24.1)	174 (23.1)	116 (15.4)
19	105 (13.9)	162 (21.5)	187 (24.8)	156 (20.7)	143 (19.0)
20	269 (35.8)	174 (23.1)	129 (17.2)	102 (13.6)	78 (10.4)
21	273 (36.4)	176 (23.4)	138 (18.4)	99 (13.2)	65 (8.7)
22	188 (25.1)	232 (31.0)	155 (20.7)	115 (15.4)	59 (7.9)
23	37 (4.9)	107 (14.2)	125 (16.6)	206 (27.4)	276 (36.8)
24	260 (34.6)	183 (24.4)	138 (18.4)	102 (13.6)	68 (9.1)
25	110 (14.6)	207 (27.5)	167 (22.2)	161 (21.4)	107 (14.2)
26	176 (23.4)	186 (24.7)	154 (20.5)	154 (20.5)	82 (10.9)
27	43 (5.7)	134 (17.8)	169 (22.5)	195 (25.9)	211 (28.1)
28	112 (14.9)	141 (18.8)	160 (21.3)	174 (23.1)	165 (21.9)
29	17 (2.3)	113 (15.0)	108 (14.4)	221 (29.4)	293 (39.0)
30	128 (17.0)	165 (21.9)	143 (19.0)	178 (23.7)	138 (18.4)
31	52 (6.9)	179 (23.8)	152 (20.2)	176 (23.4)	193 (25.7)
32	547 (72.8)	95 (12.6)	59 (7.9)	36 (4.8)	14 (1.9)
33	349 (46.4)	197 (26.2)	105 (14.0)	58 (7.7)	43 (5.7)
34	243 (32.3)	211 (28.0)	126 (16.7)	107 (14.2)	66 (8.8)

[*] Items 1, 2, 4, 8 through 10, 12, 13, 16, 25 through 29, and 34 use the following response scale: 1, not at all; 2, slightly; 3, moderately; 4, quite a bit; and 5, extremely. Items 3, 5 through 7, 11, 14, 15, 17 through 24, and 30 through 33 use the following response scale: 1, not at all; 2, slightly; 3, moderately; 4, quite a bit; 5, a great deal.

in which latent variables are multidimensional. The technical details of MH-RM are beyond the current article; however, the algorithm basically works by estimating starting values via a stochastic process—a Metropolis-Hastings sampler—and then using the values in the multivariate integrations necessary to solve the likelihood equation via the Robins-Monro root finding algorithm. Analysis between the IBS-SSS and IBS-QOL scores and model-estimated latent trait values were completed using the R package, "psych" version 1.3.12.

RESULTS

Table II lists the summary statistics for the patients. Males made up approximately 30% of the set and had a slightly higher IBS-QOL total score, on average indicating higher perceived QOL. Patients' daily means at baseline indicated moderately high values for IBS-SSS and pain and Bristol Stool Scale ratings. Of note, the current patients represent a set of patients with IBS-d because IBS-SSS scores >300 are generally thought to indicate severe disease.[18] Table III presents the frequency distributions of the item responses. Most items had balanced distribution of responses, with some exceptions. Notably, >70% of patients responded to item 32, "I fear that I won't be able to have a bowel movement," with "Not at all."

IRT and TRT Models

Model comparison statistics for the CGRM, UGRM, and TRM models are given in Table IV. The UGRM had a better fit to the data than the CGRM ($\chi^2_{34} = 1189.5$, $P < 0.0001$), indicating that individual items differ with respect to discriminating between latent trait levels. Tables V and VI give the model estimates for the CGRM and UGRM fits, respectively. The CGRM location estimates had mostly good coverage of the θ continuum (ie, on the interval −4,4), with only a few items tending toward the severe (eg, items 3 and 5) or the nonsevere (eg, item 32) ends. Generally, the UGRM location estimates covered the θ continuum well also, with a handful of items (12, 20, 21, 32, and 33) tending to be lacking in coverage of positive θ values (ie, worse QOL). Conversely, only items 5 and 23 tended to have relatively poor coverage on negative θ values.

In evaluating the differences in discrimination between the CGRM and UGRM, inspection of Table VI reveals that all but item 5 ($\beta = 0.99$, SE = 0.08) have slopes significantly higher than 1.0—the value for β in the CGRM. Item 32, an item likely more relevant to IBS with constipation, is the only item with discrimination significantly <1.0 ($\beta = 0.54$, SE = 0.09). Furthermore, there is evidence that items themselves have significantly different discrimination levels (eg, $\beta = 2.98$, SE = 0.15, for item 10 compared with $\beta = 1.75$, SE = 0.09, for item 34).

The TRM revealed an improved fit to the data than over the UGRM ($\chi^2_2 = 782.7$, $P < 0.0001$). The improved fit indicates that allowing for the covariance among the testlet items improves model estimation. Table VII gives the parameter estimates from the TRM. Generally, the parameter estimates for the TRM were similar to those for the UGRM.

Item factor loadings (λ_j) for the fitted GRM or TRM slopes[21] are given in Table VIII. From this framework, the sum of squared loadings—a measure of variance in each general factor—for the UGRM and TRM are approximately equal (16.97 vs 16.80), and both are higher than the CGRM (8.73).

As a comparison of model estimation, factor scores and information functions for θ were computed for the CGRM, UGRM, and TRMs. Figure 1 displays the

Table IV. Comparisons of the CGRM, UGRM, and TRM.

Model With Worse Fit	Model With Better Fit	LR χ^2	Δ_{df}	P
CGRM (model 1)	UGRM (model 2)	1189.549	34	<0.0001
	TRM (model 3)	1972.214	36	<0.0001
UGRM (model 2)	TRM (model 3)	782.665	2	<0.0001

CGRM = constrained graded response model; LR = likelihood ratio; TRM = testlet response model; UGRM = unconstrained graded response model.

Table V. CGRM (model 1) estimates.

		Intercepts			
Item (j)	Slope β (SE)	$\alpha_{4,5}$ (SE)	$\alpha_{3,4}$ (SE)	$\alpha_{2,3}$ (SE)	$\alpha_{1,2}$ (SE)
1	1	2.92 (0.14)	0.98 (0.10)	−0.64 (0.09)	−2.75 (0.13)
2	1	2.72 (0.13)	0.99 (0.10)	−0.34 (0.09)	−2.08 (0.12)
3	1	4.56 (0.24)	2.48 (0.12)	0.96 (0.10)	−1.07 (0.10)
4	1	1.52 (0.10)	−0.28 (0.09)	−1.84 (0.11)	−3.51 (0.16)
5	1	3.05 (0.15)	1.38 (0.10)	0.16 (0.09)	−1.51 (0.11)
6	1	1.70 (0.10)	0.09 (0.09)	−1.24 (0.10)	−2.98 (0.14)
7	1	3.24 (0.14)	1.23 (0.10)	−0.18 (0.09)	−1.80 (0.11)
8	1	2.22 (0.11)	0.56 (0.09)	−0.76 (0.10)	−2.33 (0.12)
9	1	1.50 (0.10)	−0.22 (0.09)	−1.68 (0.11)	−3.48 (0.16)
10	1	1.31 (0.10)	−0.56 (0.09)	−1.84 (0.11)	−3.55 (0.16)
11	1	3.14 (0.15)	1.70 (0.11)	0.19 (0.09)	−1.61 (0.11)
12	1	0.72 (0.09)	−0.62 (0.09)	−1.96 (0.11)	−3.48 (0.16)
13	1	1.77 (0.11)	0.03 (0.09)	−1.24 (0.10)	−2.73 (0.13)
14	1	1.32 (0.10)	−0.53 (0.09)	−1.61 (0.10)	−3.31 (0.15)
15	1	4.21 (0.20)	1.94 (0.11)	0.46 (0.09)	−1.22 (0.10)
16	1	2.61 (0.12)	0.51 (0.09)	−1.00 (0.10)	−3.06 (0.14)
17	1	1.08 (0.10)	−0.10 (0.09)	−1.05 (0.10)	−2.45 (0.12)
18	1	2.66 (0.12)	0.74 (0.09)	−0.72 (0.09)	−2.48 (0.12)
19	1	2.51 (0.12)	0.83 (0.09)	−0.63 (0.09)	−2.07 (0.11)
20	1	0.85 (0.10)	−0.46 (0.09)	−1.58 (0.10)	−2.91 (0.14)
21	1	0.85 (0.10)	−0.50 (0.09)	−1.76 (0.11)	−3.24 (0.15)
22	1	1.59 (0.10)	−0.30 (0.09)	−1.65 (0.10)	−3.35 (0.15)
23	1	3.78 (0.18)	1.93 (0.11)	0.77 (0.10)	−0.77 (0.10)
24	1	0.96 (0.10)	−0.46 (0.09)	−1.71 (0.11)	−3.21 (0.15)
25	1	2.47 (0.12)	0.52 (0.09)	−0.82 (0.10)	−2.53 (0.12)
26	1	1.71 (0.11)	0.14 (0.09)	−1.11 (0.10)	−2.90 (0.14)
27	1	3.73 (0.17)	1.68 (0.10)	0.22 (0.09)	−1.38 (0.10)
28	1	2.44 (0.12)	0.95 (0.10)	−0.34 (0.09)	−1.85 (0.11)
29	1	4.78 (0.26)	2.15 (0.11)	1.03 (0.10)	−0.70 (0.09)
30	1	2.29 (0.11)	0.66 (0.09)	−0.51 (0.09)	−2.20 (0.11)
31	1	3.49 (0.16)	1.19 (0.10)	−0.06 (0.09)	−1.56 (0.10)
32	1	−1.34 (0.10)	−2.23 (0.12)	−3.26 (0.16)	−4.79 (0.28)
33	1	0.26 (0.09)	−1.32 (0.10)	−2.56 (0.12)	−3.80 (0.17)
34	1	1.11 (0.10)	−0.52 (0.09)	−1.64 (0.10)	−3.18 (0.15)

CGRM = constrained graded response model.

densities for factor scores (ie, latent trait estimates) generated by each model. Visual differences can be seen between the CGRM and the other 2 models, with the UGRM and TRM having almost identical estimation. Figure 2 shows the information functions of the items under each of models 1 through 3. Information, either at the item or instrument level, is proportional to the reciprocal of variance and so, as a measure of precision, can be thought of as the IRT equivalent of reliability in the CTT.[21]

Table VI. UGRM (model 2) estimates.

Item (j)	Slope β_j (SE)	Intercepts			
		$\alpha_{4,5}$ (SE)	$\alpha_{3,4}$ (SE)	$\alpha_{2,3}$ (SE)	$\alpha_{1,2}$ (SE)
1	1.94 (0.11)	3.26 (0.17)	1.12 (0.11)	−0.68 (0.10)	−3.03 (0.15)
2	1.17 (0.08)	2.52 (0.13)	0.92 (0.09)	−0.30 (0.08)	−1.89 (0.11)
3	1.69 (0.11)	4.75 (0.27)	2.60 (0.14)	1.03 (0.10)	−1.09 (0.10)
4	1.55 (0.09)	1.54 (0.10)	−0.26 (0.09)	−1.82 (0.11)	−3.50 (0.18)
5	0.99 (0.08)	2.73 (0.14)	1.22 (0.09)	0.14 (0.08)	−1.33 (0.10)
6	2.48 (0.13)	2.21 (0.14)	0.16 (0.10)	−1.57 (0.12)	−3.79 (0.19)
7	2.29 (0.14)	3.96 (0.20)	1.54 (0.12)	−0.19 (0.10)	−2.15 (0.13)
8	1.28 (0.08)	2.09 (0.12)	0.53 (0.09)	−0.70 (0.09)	−2.17 (0.12)
9	2.20 (0.11)	1.81 (0.12)	−0.24 (0.09)	−1.970 (0.13)	−4.13 (0.21)
10	2.98 (0.15)	1.91 (0.13)	−0.77 (0.11)	−2.62 (0.16)	−5.13 (0.27)
11	1.25 (0.08)	2.95 (0.15)	1.59 (0.10)	0.19 (0.08)	−1.48 (0.10)
12	1.41 (0.09)	0.72 (0.09)	−0.58 (0.09)	−1.88 (0.11)	−3.36 (0.18)
13	1.84 (0.10)	1.94 (0.11)	0.06 (0.09)	−1.31 (0.11)	−2.92 (0.15)
14	1.97 (0.09)	1.50 (0.11)	−0.58 (0.10)	−1.78 (0.12)	−3.69 (0.18)
15	1.73 (0.09)	4.43 (0.23)	2.06 (0.12)	0.51 (0.09)	−1.25 (0.10)
16	2.09 (0.11)	3.04 (0.15)	0.61 (0.10)	−1.14 (0.11)	−3.52 (0.18)
17	1.54 (0.09)	1.10 (0.09)	−0.08 (0.09)	−1.03 (0.10)	−2.43 (0.14)
18	2.34 (0.10)	3.30 (0.17)	0.95 (0.11)	−0.86 (0.10)	−3.02 (0.16)
19	1.67 (0.09)	2.61 (0.13)	0.88 (0.09)	−0.63 (0.09)	−2.12 (0.13)
20	1.30 (0.08)	0.82 (0.09)	−0.42 (0.09)	−1.47 (0.10)	−2.73 (0.14)
21	1.68 (0.10)	0.90 (0.09)	−0.50 (0.09)	−1.81 (0.12)	−3.34 (0.17)
22	1.67 (0.09)	1.66 (0.11)	−0.28 (0.09)	−1.67 (0.11)	−3.44 (0.18)
23	1.13 (0.08)	3.48 (0.18)	1.76 (0.11)	0.70 (0.09)	−0.69 (0.08)
24	2.00 (0.11)	1.09 (0.10)	−0.50 (0.10)	−1.90 (0.12)	−3.61 (0.19)
25	1.77 (0.10)	2.64 (0.14)	0.58 (0.09)	−0.84 (0.10)	−2.66 (0.14)
26	1.62 (0.09)	1.76 (0.11)	0.16 (0.09)	−1.11 (0.10)	−2.93 (0.15)
27	1.87 (0.08)	4.08 (0.21)	1.87 (0.12)	0.27 (0.10)	−1.46 (0.11)
28	1.78 (0.09)	2.61 (0.14)	1.03 (0.10)	−0.34 (0.09)	−1.93 (0.12)
29	1.86 (0.10)	5.19 (0.29)	2.36 (0.13)	1.15 (0.10)	−0.72 (0.10)
30	2.60 (0.12)	3.05 (0.16)	0.93 (0.11)	−0.64 (0.11)	−2.84 (0.16)
31	1.88 (0.10)	3.82 (0.19)	1.33 (0.11)	−0.04 (0.09)	−1.67 (0.11)
32[a]	0.54 (0.09)	−1.02 (0.09)	−1.83 (0.11)	−2.73 (0.15)	−4.09 (0.27)
33	1.84 (0.10)	0.30 (0.09)	−1.41 (0.11)	−2.75 (0.15)	−4.08 (0.21)
34	1.75 (0.09)	1.20 (0.10)	−0.52 (0.09)	−1.71 (0.11)	−3.34 (0.17)

UGRM = unconstrained graded response model.

[a] Slope is less than constrained model slope.

Differential Functioning

The reduced item set was subjected to DIF analysis on sex. Table IX gives the DIF results. Seventeen items were flagged for possible DIF between males and females for the CGRM, and 9 items were likewise flagged for the UGRM. Figure 3 through B

Table VII. TRM (model 3) estimates.

	Slopes			Intercepts			
Item (j)	β_j (SE)	ζ_1 (SE)	ζ_2 (SE)	$\alpha_{4,5}$ (SE)	$\alpha_{3,4}$ (SE)	$\alpha_{2,3}$ (SE)	$\alpha_{1,2}$ (SE)
1	1.93 (0.10)			3.35 (0.16)	1.21 (0.10)	−0.61 (0.09)	−2.97 (0.15)
2	1.15 (0.07)			2.57 (0.13)	0.97 (0.09)	−0.25 (0.08)	−1.85 (0.11)
3	1.67 (0.10)			4.83 (0.26)	2.67 (0.14)	1.10 (0.09)	−1.02 (0.09)
4	1.54 (0.07)			1.61 (0.10)	−0.20 (0.08)	−1.76 (0.11)	−3.44 (0.17)
5[*]	0.97 (0.07)			2.77 (0.14)	1.26 (0.09)	0.18 (0.08)	−1.28 (0.09)
6	2.49 (0.13)			2.34 (0.13)	0.27 (0.09)	−1.48 (0.11)	−3.73 (0.19)
7	2.29 (0.12)			4.08 (0.19)	1.65 (0.11)	−0.09 (0.09)	−2.07 (0.12)
8	1.28 (0.08)			2.15 (0.11)	0.58 (0.08)	−0.65 (0.08)	−2.13 (0.12)
9	2.22 (0.08)			1.92 (0.12)	−0.15 (0.09)	−1.90 (0.12)	−4.09 (0.20)
10	2.97 (0.15)			2.06 (0.13)	−0.64 (0.11)	−2.52 (0.15)	−5.04 (0.26)
11	1.93 (0.25)	1.93 (0.25)		4.57 (0.25)	2.51 (0.15)	0.38 (0.07)	−2.15 (0.12)
12	2.71 (0.37)		2.71 (0.37)	1.48 (0.19)	−1.13 (0.16)	−3.77 (0.18)	−6.56 (0.31)
13	1.83 (0.09)			2.02 (0.12)	0.14 (0.09)	−1.24 (0.10)	−2.86 (0.15)
14	1.98 (0.10)			1.60 (0.11)	−0.50 (0.09)	−1.72 (0.11)	−3.64 (0.18)
15	1.70 (0.09)			4.49 (0.23)	2.12 (0.12)	0.58 (0.09)	−1.18 (0.09)
16	2.07 (0.09)			3.12 (0.15)	0.70 (0.09)	−1.05 (0.10)	−3.43 (0.17)
17	1.53 (0.07)			1.17 (0.09)	−0.02 (0.08)	−0.97 (0.09)	−2.37 (0.12)
18	2.33 (0.10)			3.42 (0.17)	1.06 (0.10)	−0.77 (0.10)	−2.94 (0.15)
19	1.63 (0.09)			2.66 (0.13)	0.94 (0.09)	−0.56 (0.09)	−2.04 (0.12)
20	2.24 (0.30)		2.24 (0.3)	1.51 (0.13)	−0.74 (0.18)	−2.64 (0.23)	−4.78 (0.29)
21	1.65 (0.08)			0.97 (0.09)	−0.43 (0.08)	−1.74 (0.11)	−3.27 (0.17)
22	1.64 (0.08)			1.72 (0.11)	−0.21 (0.08)	−1.60 (0.10)	−3.36 (0.17)
23	1.91 (0.26)	1.91 (0.26)		5.47 (0.30)	2.84 (0.15)	1.18 (0.10)	−1.01 (0.09)
24	1.97 (0.10)			1.17 (0.10)	−0.42 (0.09)	−1.82 (0.12)	−3.52 (0.18)
25	1.74 (0.08)			2.71 (0.13)	0.65 (0.09)	−0.77 (0.09)	−2.59 (0.13)
26	1.60 (0.07)			1.83 (0.11)	0.23 (0.09)	−1.04 (0.09)	−2.86 (0.14)
27	1.83 (0.09)			4.14 (0.20)	1.93 (0.11)	0.34 (0.09)	−1.38 (0.10)
28	2.28 (0.21)	2.28 (0.21)		4.00 (0.09)	1.67 (0.05)	−0.35 (0.07)	−2.72 (0.14)
29	1.83 (0.11)			5.25 (0.29)	2.43 (0.13)	1.23 (0.10)	−0.64 (0.09)
30	2.57 (0.13)			3.16 (0.15)	1.04 (0.10)	−0.53 (0.10)	−2.74 (0.15)
31	1.86 (0.10)			3.90 (0.19)	1.41 (0.10)	0.04 (0.09)	−1.60 (0.11)
32[†]	0.53 (0.09)			-0.99 (0.08)	−1.80 (0.11)	−2.71 (0.15)	−4.07 (0.27)
33	1.82 (0.10)			0.38 (0.09)	−1.33 (0.10)	−2.68 (0.14)	−4.00 (0.21)
34	1.75 (0.08)			1.28 (0.10)	−0.45 (0.09)	−1.64 (0.10	−3.28 (0.16)
Variances	1.00	0.98 (0.07)	1.29 (0.03)				

TRM = testlet response model.

[*] Slope is equal to constrained model slope.

[†] Slope is less than constrained model slope.

Table VIII. Comparison of factor loadings.

Item	Factor Loadings (λ_j) for θ		
	CGRM	UGRM	TRM
1	0.507	0.751	0.751
2	0.507	0.565	0.561
3	0.507	0.705	0.701
4	0.507	0.673	0.670
5	0.507	0.502	0.495
6	0.507	0.825	0.826
7	0.507	0.803	0.803
8	0.507	0.601	0.600
9	0.507	0.791	0.793
10	0.507	0.868	0.868
11	0.507	0.591	0.600
12	0.507	0.639	0.646
13	0.507	0.735	0.733
14	0.507	0.757	0.758
15	0.507	0.713	0.707
16	0.507	0.776	0.773
17	0.507	0.671	0.667
18	0.507	0.808	0.808
19	0.507	0.701	0.692
20	0.507	0.607	0.623
21	0.507	0.702	0.697
22	0.507	0.700	0.693
23	0.507	0.551	0.598
24	0.507	0.762	0.757
25	0.507	0.720	0.716
26	0.507	0.689	0.685
27	0.507	0.740	0.733
28	0.507	0.722	0.625
29	0.507	0.737	0.732
30	0.507	0.837	0.833
31	0.507	0.741	0.738
32	0.507	0.304	0.299
33	0.507	0.734	0.730
34	0.507	0.718	0.716
Sums of squared loadings	8.725	16.965	16.799

CGRM = constrained graded response model; LR = likelihood ratio; TRM = testlet response model; UGRM = unconstrained graded response model.

also graphically represent the effect of patient sex on estimation of the latent trait for all 3 models explored.

Associations Between Measures

To determine how well the latent variable estimation of the CGRM, UGRM, and TRMs relate back to traditional, true-score theory, outcome measures, bivariate Spearman correlation coefficients were calculated between each model estimate of θ and the IBS-SSS and 34-item IBS-QOL score, both of which have been used previously to measure IBS disease states.[6,18] Of note, for the IBS-SSS, a higher score is indicative of more severe symptoms, higher IBS-QOL total scores are indicative of better QOL, and higher scores on the latent estimate scales are indicative of worse QOL.

Figure 6 illustrates the associations among the different scales. Interestingly, the IBS-QOL questionnaire correlates extremely with the different model-based estimates of θ. These results suggest that the IBS-QOL is a good surrogate for the latent variable, perceived QOL in patients with IBS-d. Furthermore, the strength of the correlations indicates that the associations are linear, thereby suggesting that the IBS-QOL questionnaire is a good measure for patients over a wide range of health-related QOL status. The results are supported by a smaller but still strong correlation with the IBS-SSS for the IBS-QOL questionnaire and the model estimates.

DISCUSSION

Three IRT models were fit to baseline data from a Phase II intervention trial, and increasing complexity of the models generally improved fit. These results are not unexpected because most data sets reveal increasingly better fit with more complicated models. It is also not surprising that freely estimating slopes of items improves model fit. The UGRM model fit here actually closely resembles the previous factor analyses that have been conducted in that the items are allowed to have different contributions to the estimation of the latent trait/general factor (ie, perceived QOL [θ]).[6,9]

Unidimensional IRT models assume that there is a single underlying trait for estimation of θ, and they further assume local independence of items. Our TRM model assessed whether there were dependencies between 2 sets of items that related, conceptually, to

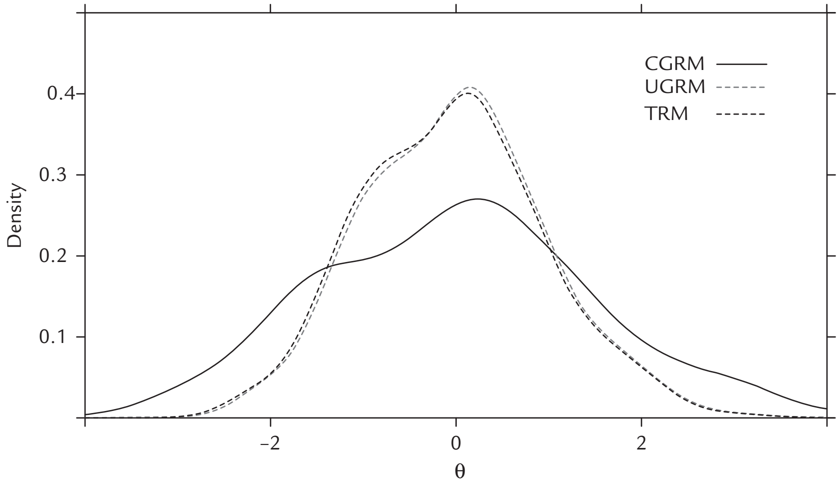

Figure 1. Comparison of estimated latent trait values (θ) for the constrained graded response model (CGRM), the unconstrained GRM (UGRM), and the testlet response model (TRM).

eating and sexual habits in relation to patients' IBS-d. Our results indicate that although there are numeric differences between the TRM and UGRM fits, estimation of perceived QOL is virtually the same between these 2 approaches. Because IRT and TRT models estimate latent traits on a continuum instead of as a single true score, the information can vary over the values of θ. A useful metric for determining information is the area under the information curve, which is presented in the legend for Figure 2. An area under the information curve gives a single metric of model precision over the range of θ, with higher values generally indicating that θ is measured more accurately. Visual inspection of the information

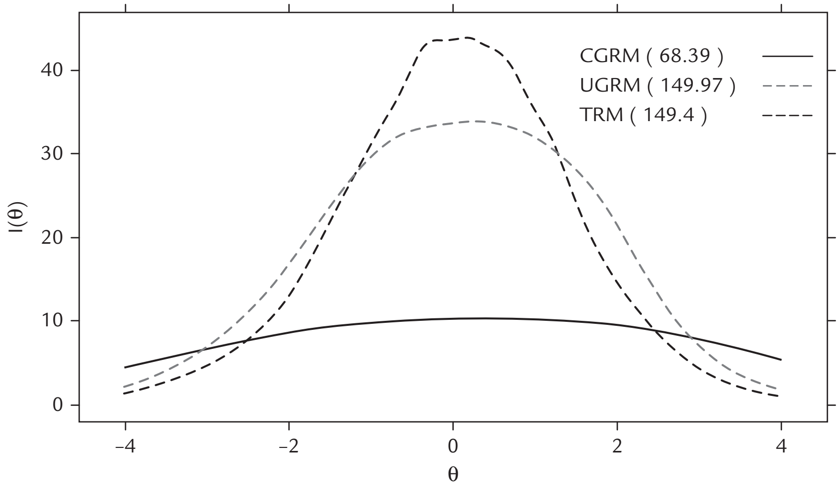

Figure 2. Comparison of information functions, I(θ), for the constrained graded response model (CGRM), the unconstrained GRM (UGRM), and the testlet response model (TRM), with total information as determined by the area under each information curve.

Table IX. Items and testlets that revealed possible differential functioning due to sex.

	CGRM		UGRM	
Item or Testlet Excluded	LR χ^2	$\Pr(\chi^2)$[a]	LR χ^2	$\Pr(\chi^2)$[a]
1. I feel helpless because of my bowel problems.	60.89	<0.001		
2. I am embarrassed by the smell caused by my bowel problems.	128.37	<0.001	14.11	0.0293
4. I feel vulnerable to other illnesses because of my bowel problems.	55.38	<0.0001		
5. I feel fat because of my bowel problems.	91.42	<0.0001	36.89	<0.0001
6. I feel like I'm losing control of my life because of my bowel problems.	35.97	<0.0001		
11. I have to watch the amount of food I eat because of my bowel problems.	35.63	<0.0001		
13. I feel angry that I have bowel problems.			26.58	<0.0001
14. I feel like I irritate others because of my bowel problems.			19.72	0.0018
15. I worry that my bowel problems will get worse.	35.67	<0.0001		
16. I feel irritable because of my bowel problems.			17.26	0.0061
17. I worry that people think I exaggerate my bowel problems.	27.39	<0.0001		
19. I have to avoid stressful situations because of my bowel problems.	42.54	<0.0001		
21. My bowel problems limit what I can wear.	81.17	<0.0001	28.80	<0.0001
22. I have to avoid strenuous activity because of my bowel problems.	67.40	<0.0001		
23. I have to watch the kind of food I eat because of my bowel problems.	29.31	<0.0001		
25. I feel sluggish because of my bowel problems.	127.57	<0.0001		
26. I feel unclean because of my bowel problems.	40.61	<0.0001		
27. Long trips are difficult for me because of my bowel problems.	47.36	<0.0001		
28. I feel frustrated that I cannot eat when I want because of my bowel problems.	47.48	<0.0001		
29. It is important to be near a toile t because of my bowel problems.			16.98	0.0070
31. I worry about losing control of my bowels.	53.08	<0.0001		
33. My bowel problems are affecting my closest relationships.			39.02	<0.0001
34. I feel that no one understands my bowel problems.			16.49	0.0089
I. Items 11, 23, and 28	<1.00	>0.9999		
II. Items 12 and 20	<1.00	>0.9999		

CGRM = constrained graded response model; LR = likelihood ratio; Pr = probability; UGRM = unconstrained graded response model.

[a] Displayed *P* values are adjusted using the Bonferroni procedure.

functions reveals that both the UGRM and TRM hold more information than the CGRM. The total information values for the UGRM and TRM (149.97 and 149.40, respectively) are higher than that for the CGRM (68.39). Interestingly, although the UGRM and TRM information levels are virtually equal, their information functions differ. The UGRM has a wider and flatter information function, whereas the TRM information function has a much more distinctive peak. One interpretation of this pattern is that the TRM helps discriminate finely around the mean at the expense of discrimination away from the mean, whereas the UGRM has a better overall coverage of the θ continuum. The only notable difference in information estimation was that the TRM model appears be superior to the UGRM around the sample mean; however, the TRM did not maintain this higher level of discrimination over the entire θ continuum. One interpretation of these results is that the TRM may provide added discrimination between

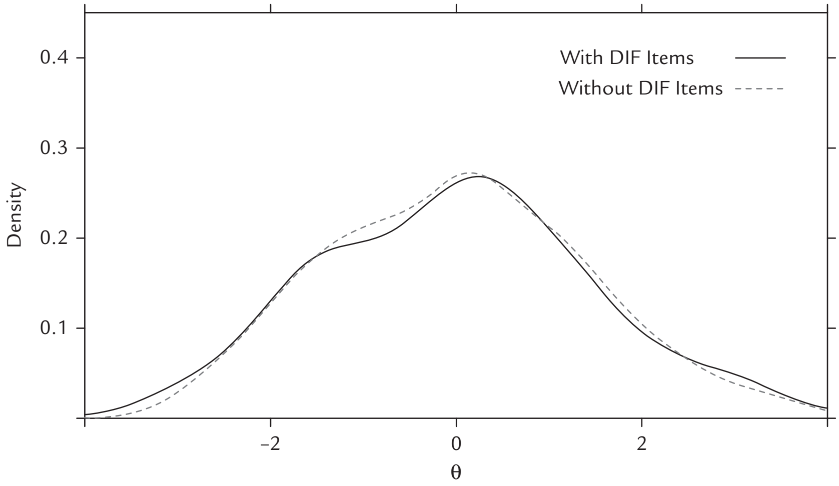

Figure 3. Effect of items with possible differential item functioning (DIF) due to sex on constrained graded response model estimates of the latent trait (θ).

IBS-d and other IBS subtypes in a mixed analysis set—an interesting future potential research area.

Likewise, potential DIF associated with patient sex revealed numeric differences on certain items but did not seem to affect overall estimates of perceived QOL—as measured by θ—for any of the GRM or TRM approaches. With regard to the UGRM, despite 9 items revealing potential DIF at the Bonferroni-adjusted $\alpha = .05$ level, the effect on the overall density of the latent trait is negligible (Figure 4), as evidenced by the overlap in densities between estimates from the full item set vs the set of 25 items not flagged for potential DIF. Neither of the 2 testlets exhibited potential DIF by the LR criterion. There is of interest when exploring the potential DIF of the IBS-QOL items among racial groups. For the current data, however, nonwhite patients only totaled 107, a number deemed currently too small to evaluate DIF. Evaluating potential DIF in racial

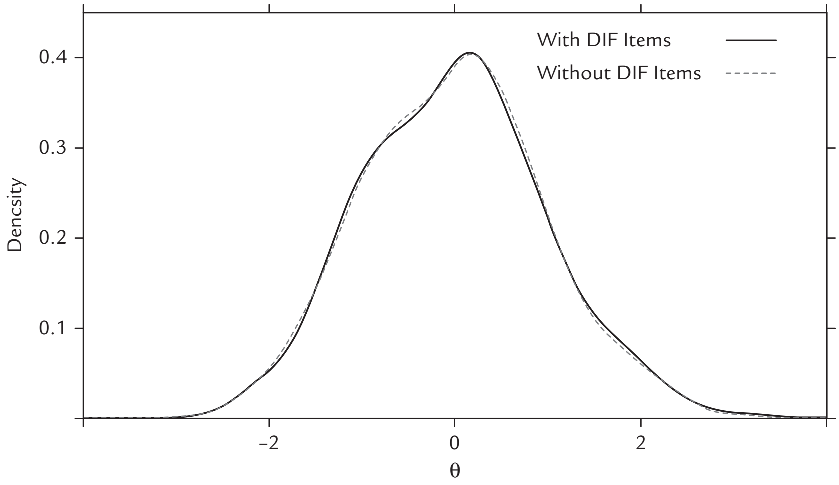

Figure 4. Effect of items with possible differential item functioning (DIF) due to sex on unconstrained graded response model estimates of the latent trait (θ).

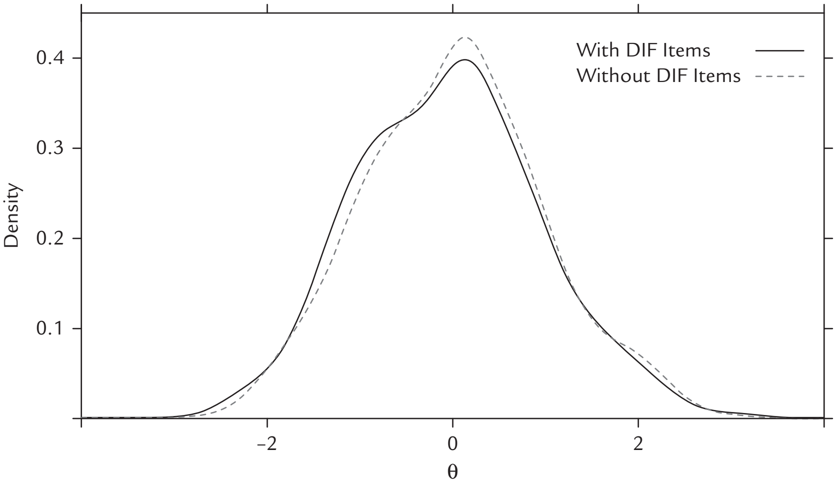

Figure 5. Effect of items with possible differential item functioning (DIF) due to sex on the testlet response model estimates of the latent trait (θ).

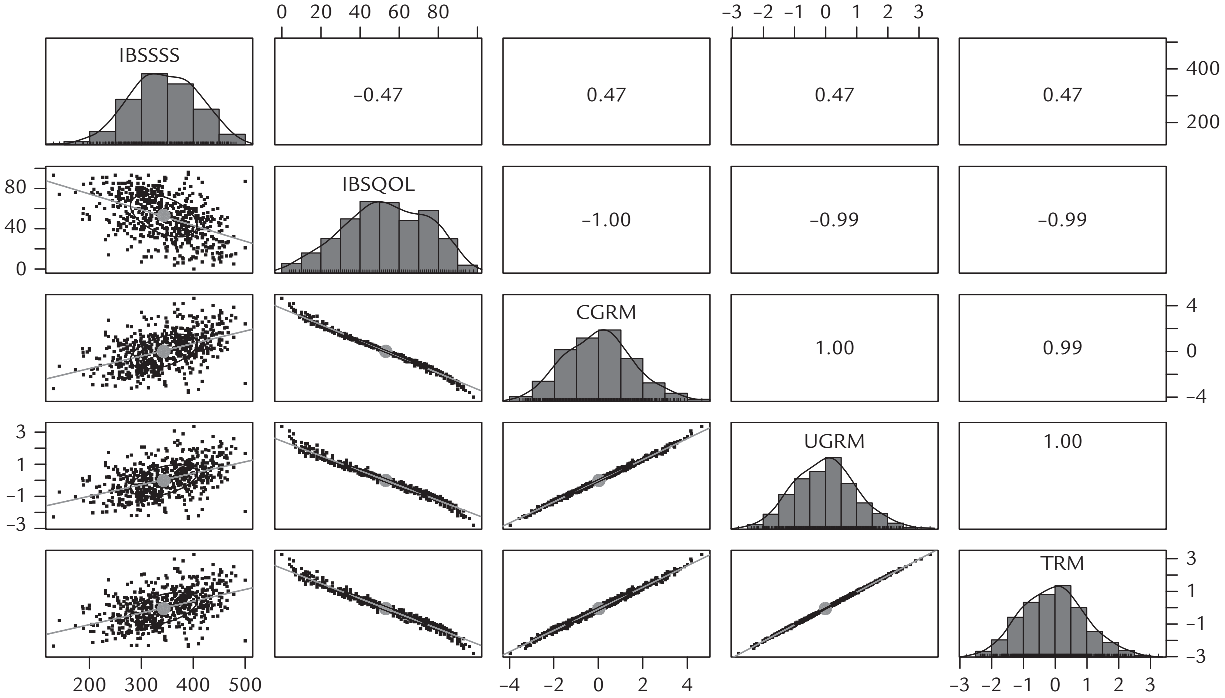

Figure 6. Scatterplot matrix of the IBS-SSS, 34-item Irritable Bowel Syndrome Quality of Life (IBS-QOL) total score with the latent trait (θ) estimates from the constrained graded response model (CGRM), the unconstrained GRM (UGRM), and the testlet response model (TRM) vs the IBS-QOL total score for 32 items. The lower triangle represents scatterplots with regression lines overlaid, the diagonal depicts the data densities of the scores, and the upper triangle displays the Spearman correlation coefficients among the scores.

and ethnic groups is a possibility for future research and may be possible when currently enrolling large-scale Phase III studies are completed.

Despite the observed gains in model fit with TRM models, comparison of the estimated perceived QOL from the CGRM, UGRM, and TRM approaches were not only found to be consistent with the IBS-SSS but also correlated at an extremely high level with the original scoring of the IBS-QOL on the 34-item set ($|r| \geq 0.99$ for all parwise comparisons of IBS-QOL and latent trait estimates). These results underscore the consistency in the items for the data set analyzed in terms of measuring latent perceived QOL related to IBS-d. Interestingly, despite the fact that the IBS-SSS is more of a symptom inventory and the IBS-QOL questionnaire is intended to measure perceived QOL, the Spearman correlation of 0.47 indicates that symptom severity and perceived QOL are related at a moderate to high level.[22]

Our results support previous IBS-QOL validations in many ways. Despite increasing item fits with complexity of the models, one can still conclude that the IBS-QOL questionnaire is measuring a unidimensional construct. When viewing the model slope results from a factor analytic perspective, the results indicate not only that free estimation of item slopes help measurement of perceived QOL in patients with IBS-d but also that the models support the idea of approximate unidimensionality of θ because the model loadings are all positive and high by usual evaluation standards for factor analysis. Furthermore, comparison among the IBS-SSS, IBS-QOL total score, and the factor scores from CGRM, UGRM, and TRM models yielded strikingly consistent results, which further buttressed the position that the IBS-QOL questionnaire is measuring perceived QOL in IBS-d patients and doing so well (**Figure 5**).

CONCLUSION

Irritable bowel syndrome is a condition in which severity is measured by patient reports of their physical and psychological symptoms. The IBS-QOL questionnaire is a psychometrically sound instrument for measuring perceived QOL in both male and female patients with IBS-d. Although there is evidence for local dependencies among individual items of the IBS-QOL questionnaire, these local dependencies do not appear to detrimentally influence perceived QOL measurement, and we have established that the original scoring is a viable surrogate for perceived QOL. Adding model complexity within a structure that corresponded with theory around the IBS-QOL measurement model revealed superior results; however, it appears that in the case of the IBS-QOL, the added model complexity simply added to our understanding of the baseline measurement properties of the instrument and did not offer much in the way of improving the scoring of the IBS-QOL. Rather, the original scoring of the IBS-QOL stands up well with estimates from more complex latent variable modeling techniques. The IBS-QOL was developed under a needs-based conceptual model.[6] This conceptual model has held up well to quantitative scrutiny that further underscores its validity for measuring perceived QOL.

ACKNOWLEDGMENTS

We thank Mike Davenport and one anonymous referee for constructive comments on previous versions of the manuscript. Randi Gress also aided in the preparation of the manuscript.

Dr. Andrae aided in study design, conceived the data analysis, conducted and interpreted analyses, and drafted the manuscript. Dr. Covington aided in study design, data interpretation, and drafting of the manuscript. Dr. Patrick aided in data interpretation and manuscript drafting.

CONFLICTS OF INTEREST

David A. Andrae and Paul S. Covington are employees and stockholders of Furiex Pharmaceuticals. Donald L. Patrick is a paid consultant to Furiex Pharmaceuticals

REFERENCES

1. Lovell R, Ford A. Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. *Clin Gastroenterol Hepatol*. 2012;10:712–721.
2. Drossman D. The functional gastrointestinal disorders and the Rome III process. *Gastroenterology*. 2006;130:1377–1390.
3. Drossman D. *Rome III The Functional GI Disorders*. 3rd ed. Lawrence, KS: Allen Press; 2006.
4. Brandt L, Chey W, Foxx-Orenstein A. An evidence-based position statement on the management of irritable bowel syndrome. *Am J Gastroenterol*. 2009;104:1–35.
5. Dove L, Lembo A, Randall C, et al. Eluxadoline benefits patients with irritable bowel syndrome with diarrhea in a phase 2 study. *Gastroenterology*. 2013;145:329–338.
6. Patrick D, Drossman D, Frederick I, Dicesare J, Puder K. Quality of life in persons with irritable bowel syndrome:

development and validation of a new measure. *Dig Dis Sci*. 1998; 43:400–411.

7. Drossman D, Patrick D, Whitehead W, et al. Further validation of the IBS-QOL: a disease-specific quality-of-life questionnaire. *Am J Gastroenterol*. 2000;95:999–1007.
8. Drossman D, Morris CB, Hu Y, et al. Characterization of health related quality of life (HRQOL) for patients with functional bowel disorder (FBD) and its response to treatment. *Am J Gastroenterol*. 2007;102:1442–1453.
9. Andrae DA, Patrick DL, Drossman DA, Covington PS. Evaluation of the Irritable Bowel Syndrome Quality of Life (IBS-QOL) questionnaire in diarrheal predominant irritable bowel syndrome patients. *Health Qual Life Outcomes*. 2013;11:208.
10. Samejima F. *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Richmond, VA: Psychometric Society; 1969.
11. Hambleton R, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff Publishing; 1985.
12. Masters GN, Wright BD. The partial credit model. In: van der Linden WJ, Hambleton RK, eds. *The Handbook of Modern Item Response Theory*. New York, NY: Springer; 1997:101–122.
13. Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika*. 1986;51:567–577.
14. Wang X, Bradlow ET, Wainer H. *Testlet Response Theory and Its Applications*. Cambridge, UK: Cambridge University Press; 2007.
15. Wainer H, Wang X. *Using a New Statistical Model for Testlets to Score TOEFL*. Princeton, NJ: ETS; 2001.
16. Wang X, Bradlow ET, Wainer H. *A General Bayesian Model for Testlets: Theory and Applications*. Princeton, NJ: ETS; 2002.
17. Holzinger K, Swineford F. The bi-factor method. *Psychometrika*. 1937;2: 41–54.
18. Francis C, Morris J, Whorwell P. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. *Aliment Pharmacol Ther*. 1997;11:395–402.
19. Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*. 2010;75:33–57.
20. Cai L. A two-tier full-information item factor analysis model with applications. *Psychometrika*. 2010; 75:581–612.
21. Baker FB. *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker; 1992.
22. Cohen J. A power primer. *Psychol Bull*. 1992;112:155–159.

Address correspondence to: David A. Andrae, PhD, Furiex Pharmaceuticals Inc, 3900 Paramount Pkwy, Suite 150, Morrisville, NC 27560. E-mail: david.andrae@furiex.com